# Personality Trait Differences between Young and Middle-Aged Adults: Measurement Artifacts or Actual Trends?

Nye, Christopher D ; Allemand, Mathias ; Gosling, Samuel D ; Potter, Jeff ; Roberts, Brent W

DOI: https://doi.org/10.1111/jopy.12173

Running head: MEASUREMENT EQUIVALENCE ACROSS AGES

# Personality Trait Differences between Young and Middle-Aged Adults: Measurement Artifacts or Actual Trends?

Christopher D. Nye
Michigan State University

Mathias Allemand
University of Zurich

Samuel D. Gosling
University of Texas at Austin
University of Melbourne

Jeff Potter
Atof, Inc., Cambridge, Massachusetts

Brent W. Roberts
University of Illinois, Urbana-Champaign

Correspondence concerning this article should be addressed to Christopher D. Nye, Department of Psychology, Michigan State University, 316 Physics Rd., East Lansing, MI 48824. Phone number (517) 355-3408. Electronic mail may be sent to nyechris@msu.edu.

An earlier version of this paper was presented at the annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA, April, 2012.

Abstract

**Objective**

A growing body of research demonstrates that older individuals tend to score differently on personality measures than younger adults. However, recent research using item response theory (IRT) has questioned these findings, suggesting that apparent age differences in personality traits merely reflect artifacts of the response process rather than true differences in the latent constructs. Conversely, other studies have found the opposite—age differences appear to be true differences rather than response artifacts. Given these contradictory findings, the goal of the present study was to examine the measurement equivalence of personality ratings drawn from large groups of young and middle-aged adults to 1) examine whether age differences in personality traits could be completely explained by measurement nonequivalence, and 2) to illustrate the comparability of IRT and CFA approaches to testing equivalence in this context.

**Method**

Self-ratings of personality traits were analyzed in two groups of Internet respondents aged 20 and 50 (n = 15,726 in each age group).

**Results**

Measurement nonequivalence across these groups was negligible. The effect sizes of the mean differences due to nonequivalence ranged from -.16 to .15.

**Conclusions**

Results indicate that personality trait differences across age groups reflect actual differences rather than merely response artifacts.

**Keywords**: Age differences in personality; Measurement Equivalence; Differential Item Functioning

# Personality Trait Differences between Young and Middle-Aged Adults: Measurement Artifacts or Actual Trends?

Research on age differences in personality traits has coalesced into a strikingly coherent picture. Starting several decades ago, cross-sectional research has consistently shown that middle-aged individuals tend to score higher than young adults do on agreeableness and conscientiousness and lower on extraversion, neuroticism, and openness (Costa & McCrae, 1988). Recent studies tracking age differences across larger samples and different countries have shown the same trends (Jackson et al., 2009; Labouvie-Vief, Diehl, Tarnowski, & Shen, 2000; Donnellan & Lucas, 2008; Lucas & Donnellan, 2009; Soto, John, Gosling, & Potter, 2011; Srivastava, John, Gosling, & Potter, 2003). In addition, it is increasingly difficult to argue that this pattern is the result of cohort differences because it continues to be replicated with younger and younger cohorts (Smits, Dolan, Vorst, Wicherts, & Timmerman, 2011).

Similar results can be found in longitudinal studies, mostly drawn from Western cultures. A meta-analysis of 92 longitudinal studies with various samples that covered the life course from ages 10 to 101 (Roberts, Walton, & Viechtbauer, 2006) found that most samples demonstrated increases in conscientiousness and agreeableness and decreases in neuroticism. Most interestingly, subsequent longitudinal studies across various age groups and time spans have found similar overall patterns of longitudinal change in personality traits (e.g., Allemand, Zimprich, & Hertzog, 2007; Bleidorn, Kandler, Riemann, Angleitner, & Spinath, 2009; Donnellan, Conger, & Burzette, 2007; Specht, Egloff, & Schmukle, 2011). The findings in cross-sectional and longitudinal studies are so consistent that these differences were codified into a principle of personality development labeled the maturity principle (Roberts & Wood, 2006).

People who are more mature tend to be more agreeable, conscientious, emotionally stable, and socially dominant (i.e., a facet of extraversion).

Recently, an alternative to the maturity hypothesis was proposed. This idea, which we refer to as the artifact hypothesis, holds that the age differences found across numerous cross-sectional and longitudinal studies are an artifact of the way older and younger respondents use the rating scales found in personality inventories (Tackett, Balsis, Oltmanns, & Krueger, 2009). Using an item-response theory (IRT) analysis, the Neuroticism scale showed differential test functioning (DTF) across age groups. DTF occurs when the psychological meaning of the scale scores differ across groups or over time. For example, older individuals may understand the meaning of an item assessing neuroticism in a qualitatively different way than their younger counterparts. Several factors can cause DTF but the result is that the personality scores across age groups cannot be justifiably compared because responses in each group will be on a different metric. These differences in the response process may be important for understanding personality within each age group but they become problematic when comparisons are made across age groups. Consequently, if the artifact hypothesis is true, then the conclusion that personality trait scores vary with age could be confounded with differences in the measurement process.

Other studies of measurement invariance across groups have reached different conclusions. For example, Allemand et al. (2007), found no evidence of measurement nonequivalence across middle-aged and older adults using a German version of the NEO-FFI (Costa & McCrae, 1992). Subsequent studies have shown similar patterns of equivalence in cross-sectional studies of Big Five personality trait differences in the Netherlands (Allemand, Zimprich, & Hendriks, 2008), the United States (Jackson et al., 2009), in older populations

(Small, Hertzog, Hultsch, & Dixon, 2003), and for different types of personality measures, such as self-concept clarity (Lodi-Smith & Roberts, 2010).

Given its potential to undermine the conclusion that personality traits differ across young and middle-aged adults, the artifact hypothesis deserves further attention. Two different analytic approaches have been used in past studies to examine this issue. In addition to the IRT approach used by Tackett et al. (2009), other studies have relied on an alternative technique, known as mean and covariance structures analysis (MACS; Little, 1997). A MACS approach tests the measurement equivalence of scales using a confirmatory factor analytic approach estimated within a structural equation modeling (SEM) framework. Tackett et al. identified nonequivalence using IRT whereas studies using the MACS approach have generally found equivalence in longitudinal and cross-sectional studies (e.g., Allemand et al., 2007; Jackson et al., 2009, Small et al., 2003). Despite this inconsistency, past research has established the intrinsic commonalities of IRT and confirmatory factor analytic approaches to detecting differential item functioning (DIF; Stark, Chernyshenko, & Drasgow, 2006). Given that IRT approaches have not been formally compared to MACS approaches in the literature on age differences in personality traits, we examined DIF and DTF in personality ratings drawn from large groups of young and middle-aged adults to 1) test the artifact hypothesis, and 2) to illustrate the comparability of IRT and MACS approaches to testing measurement equivalence in this context.

We also examined the implications of DIF for understanding personality trait differences across age groups. If DIF is identified, recent advances in IRT (Stark, Chernyshenko, & Drasgow, 2004) and CFA (Nye & Drasgow, 2011a) can be used to control for DIF and calculate the true differences between groups or over time. As such, although DIF may confound mean-level comparisons of personality trait scores across groups, it does not preclude the possibility

that true differences also exist. The ability to partial out the effects of DIF and estimate the effect size of true differences across groups is a recent development in the literature on measurement nonequivalence and, therefore, has not yet been applied to understand the implications of DIF for studies examining age differences in personality traits.

In the following section we describe IRT and MACS methodologies to lay the foundation for our analyses and provide a conceptual comparison of these approaches. We then test the artifact hypothesis by using both IRT and MACS analyses to examine measurement equivalence and its implications for age differences in Big Five personality trait ratings drawn from a large Internet sample.

*IRT DIF Analyses*

IRT describes an individual's responses to a personality assessment as a function of the characteristics of the items and his or her standing on the latent trait. Within the IRT framework, several models are available for defining this relationship between items and traits. One of the more popular models in personality research is Samejima's Graded Response (SGR; Samejima, 1969) model. This model was used by Tackett et al. (2009) to examine DTF across age groups and has been shown to be the best IRT model for modeling Likert-type personality data (Maydeu-Olivares, 2005). The SGR model represents the response options for a particular item as ordered categories and is described in more detail in the Appendix.

The SGR model is a unidimensional model that describes the relationship between a single latent trait and item responses. As such, this model can only be used on unidimensional constructs. This limitation is potentially problematic for personality scales that are often multidimensional (Maydeau-Olivares, 2005; Nye, Roberts, Saucier, & Zhou, 2008). Fortunately, multidimensional IRT models are also available (Reckase, 2009) and a multidimensional SGR

model can be estimated (Forero & Maydeu-Olivares, 2009). These multidimensional models are complicated so we do not provide a comprehensive description of them here. Instead, interested readers are referred to Reckase (2009) for a comprehensive resource on this topic.

*Evaluating DIF with IRT*

It is important to note that the benefits of IRT for DIF detection are only realized when an appropriate model is fit to the data. Therefore, chi-square fit statistics for both single items and pairs of items can be used to investigate the extent of model fit (Drasgow, Levine, Tsien, Williams, & Mead, 1995)[1]. After identifying an appropriate IRT model, measurement equivalence is assessed using IRT DIF techniques. DIF occurs when individuals with the same score on the latent trait, but sampled from different subpopulations, have different expected scores on an item (Drasgow, 1984). In other words, DIF can be represented as differences between the predicted responses for individuals in the reference and focal groups. In IRT terminology, the reference group is the comparison or baseline group and the focal group is the sample being examined for DIF. In practice, the reference and focal groups are typically the majority and minority groups, respectively.

In the present study, we used the likelihood ratio (LR) method (Thissen, Steinberg, & Wainer, 1993) to identify DIF. The LR method involves comparing the fit (i.e., likelihood) of a model with item parameters constrained to be equivalent across groups to the fit of an unconstrained model where all of the parameters are freely estimated in each group. The difference in the fit of these two models follows a chi-square distribution with degrees of freedom equal to the difference in the degrees of freedom for each model. As such, significance tests can be conducted and DIF is identified if the differences between these models are statistically significant.

[1] See the Appendix for a more detailed description of methods for examining IRT model fit.

With IRT models, DIF can also be aggregated to the test level to examine differential test functioning (DTF). DTF is the sum of differential functioning at the item level so biases in the opposite directions can cancel one another out at the test level. Thus, although DIF may be present in all of the items, significant DTF may not be. In this case, non-significant DTF suggests that test-score differences across groups are not attributable to DIF. To facilitate the interpretation of DTF, Stark, Chernyshenko, and Drasgow (2004) developed an effect size measure for these analyses[2]. This measure is a standardized effect size and, therefore, can be interpreted using Cohen's (1988) guidelines. Specifically, effect sizes between .20 and .50 are considered small, between .50 and .80 are considered medium, and greater than .80 are viewed as large. Anything less than .20 is considered negligible.

*MACS Analyses*

Similar to IRT, CFA models describe an individual's response to an item as a function of the characteristics of the item and his or her standing on the latent trait. However, in contrast to the non-linear IRT functions, the CFA model defines a linear relationship between the latent trait and item responses. Despite the computational differences between IRT and CFA models, several authors have demonstrated the mathematical equivalence[3] of these techniques and suggested that they are two alternative ways of describing the same model (Forero & Maydeu-Olivares, 2009; McDonald, 1999; Takane & de Leeuw, 1987). In other words, both approaches can be used interchangeably to analyze the same sets of data.

---

[2] Stark et al. (2004) defined an effect size as the magnitude of the differences between the test characteristic curves (TCCs; the sum of the ICCs) across the latent trait distribution. When this value is divided by the standard deviation of the focal group, the effect size measure is in a standardized metric similar to Cohen's *d*

[3] A full discussion of the mathematical proof of the equivalence of these two models is beyond the scope of the present study so interested readers are referred to McDonald (1999) or Takane and de Leeuw (1987) for more quantitative discussions of the similarities and differences between IRT and CFA.

As with the IRT approach, an appropriate CFA model must be identified before nonequivalence can be examined. Traditionally, a chi-square statistic has been used to examine model fit in the CFA framework. However, this index is affected by both the sample size and model complexity (Byrne, 1998; Lei & Lomax, 2005; Meade, Johnson, & Braddy, 2008). As such, alternative goodness-of-fit indices that correct for these factors have been developed and are widely used. Some of the more popular indices include the Root Mean Square Error of Approximation (RMSEA), the comparative fit index (CFI), the non-normed fit index (NNFI), and the standardized root mean square residual (SRMR)[4].

Once a good fitting CFA model is identified, measurement equivalence is tested by estimating a series of models that test for differences in the model parameters across groups. The first model (configural invariance model) examines whether the factor structure is the same in each group. If this is the case, then the factor loadings (metric invariance model) and intercepts (scalar invariance model) are constrained to be equal across groups and DIF is identified by evaluating the changes in model fit with each successive constraint. In the present study, we focus on $\Delta CFI > .002$ which has been shown to be the best indicator of nonequivalence across groups (Cheung & Renvold, 2002; Meade et al., 2008). However, we also examine the $\chi^2$ difference tests for comparison with the IRT results. More detailed information about the sequential tests for nonequivalence in the MACS approach and the criteria for identifying DIF are provided in the Appendix.

*Consequences of DIF*

Whether DIF is identified using the IRT or CFA approach, the consequences of DIF will be the same. DIF in a measure can affect both the mean and variance of a scale as well as the

[4] A full description of these indices is beyond the scope of this paper but additional information can be found in Bollen (1989) or Hu and Bentler (1999).

correlation of that scale with other variables (Nye, 2011; Nye & Drasgow, 2011a). As a result, it is possible that DIF could result in the mean-level differences that have been observed in personality traits across age groups, as predicted by the artifact hypothesis (Tackett et al., 2009).

However, it is important to note that the presence of DIF does not preclude the existence of true mean-level differences. In fact, observed differences between groups are the sum of both DIF and impact, where impact refers to true mean-level differences. In other words, the mean differences that have been observed across age groups may be a combination of both true mean differences and DIF. Using CFA, Nye and Drasgow (2011a) developed a methodology for differentiating between these two sources of observed differences. These authors showed that within a measure of the Big Five personality traits (i.e., the Mini-Marker Scale; Saucier, 1994), the percentage of the observed differences between cultures that could be attributed to DIF was as low as 18%--indicating that significant differences remained even after the effects of nonequivalence were accounted for. As described above, similar techniques for differentiating DTF from impact are also available in IRT (Stark et al., 2004).

As previous studies suggest, identifying DIF may not be enough to draw conclusions about the implications of observed differences over time or between age groups. Therefore, we examined measurement equivalence across age groups and explored the differential effects of DIF and impact on observed differences. This approach allowed us to demonstrate the implications of DIF for understanding age differences in personality traits.

## Methods

### *Sample*

The overall sample for this study consisted of approximately 2.4 million respondents to an online survey. Of this sample, 57.5% of respondents were female and 58.2% were Caucasian.

The mean age in the overall sample was 25.48. The data come from visitors to outofservice.com, which hosts a non-commercial, advertisement-free website containing a variety of personality measures (for details see Gosling, Vazire, Srivastava, & John, 2004; Rentfrow, Gosling, & Potter, 2008; Soto et al., 2011; Srivastava et al., 2003). Respondents could learn about the project through several channels, including search engines or links on such websites as www.socialpsychology.org. After submitting their responses, participants received customized feedback about their personalities. Analyses based on this data set have shown that, compared to conventional samples, Internet samples are more diverse with respect to gender, socioeconomic status, geographic region, and age (Gosling et al., 2004). Moreover, Internet findings generalize across presentation formats, are not adversely affected by non-serious or repeat responders, and are generally consistent with findings from traditional methods.

Previous research has shown that the majority of personality trait change occurs between the ages of 20 and 50 (Roberts et al., 2006), with most gains being made by age 50 (Soto et al., 2011). Consequently, the period between these two age groups represents the time period during which the most substantial changes in personality traits occur. If measurement nonequivalence exists, it is likely to be observed between the age groups with the largest differences. Therefore, analyses for the present study were conducted on two groups within this broader sample—the reference group consisted of the sample of respondents that were 20 years old and the focal group was comprised of individuals that were 50 years old.

The number of individuals at age 20 (N = 142,605) was substantially larger than the 50-year-old sample (N = 15,726). Previous research has shown that differences in the sample sizes for the reference and focal groups can result in high Type I error rates (i.e., a high number of items falsely identified as nonequivalent) when using MACS analyses (Gonzalez-Roma,

Hernandez, & Gomez-Benito, 2006). Therefore, a random sub-sample of 15,726 individuals was selected from the 20-year-old group so that the sample size would be the same for both groups.

*Measure*

Personality traits were assessed using the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991). The BFI is widely used and was developed as a relatively short assessment of the Big Five traits which include extraversion, agreeableness, conscientiousness, neuroticism, and openness. This measure assesses the five traits using 44 phrase items that were based on the trait adjectives that are prototypical markers of each trait. Responses to each item are provided on a 5-point scale ranging from 1 (*disagree strongly*) to 5 (*agree strongly*). Alpha reliabilities for the BFI scales in the full sample ranged from .77 to .85 and test-retest reliabilities are typically between .80 and .90 (Benet-Martínez & John, 1998).

*Analyses*

As described above, the benefits of both IRT and CFA are only obtained when the model accurately describes the response process. Therefore, we used adjusted $\chi^2$/df ratios to evaluate IRT model fit. Here, the $\chi^2$/df ratio is statistically adjusted[5] to a smaller sample size (we used N = 500 in this study) and ratios greater than 3 indicate misfit (Chernyshenko et al., 2007). In addition, we used traditional fit indices to evaluate CFA model fit including the chi-square statistic, RMSEA, CFI, NNFI, and SRMR. Additional information about evaluating the fit of IRT and CFA models is provided in the Appendix.

Before DIF or DTF can be assessed using IRT, item parameters and theta estimates for both samples must be equated to the same scale. IRT parameters are theoretically invariant across groups and items but estimates of their true values will be affected by the sample characteristics (see Embretson & Reise, 2000 for a simulated example). For example, a

[5] See the Appendix for the equation that was used for this adjustment.

neuroticism item may have a higher estimated *b*-parameter in a low neuroticism group than in a high neuroticism group. Consequently, parameters in the reference and focal groups must be put on a common metric. In the present study, this was achieved by estimating the IRT parameters in each group simultaneously and using anchor items to equate the metrics. Here, the parameters for the anchor items were constrained to be equivalent across groups, which places the parameter estimates in both samples on a common metric.

Similar to DIF analysis, CFA MACS analysis requires unbiased items to set the scale of the latent factors. The latent factors in these CFA models are unobservable and do not have an inherent scale so they must be given one by applying constraints to the model. The most popular method of doing this is to constrain the factor loading of one of the items, referred to as the referent item, to 1.00. In addition, the intercept of this same item can be constrained to 0 to set the mean of the latent factor.

After identifying and constraining the referent and anchor items (hereafter referred to only as referent items for both CFA and IRT methods), a free-baseline approach was used to examine non-equivalence with both IRT and CFA. With this approach, an unconstrained model, with only the parameters for the referent item constrained to be equivalent across groups[6], was used as the baseline for comparison. Then the parameters for a single item at a time were constrained to be equivalent across groups and DIF was identified by decreases in the CFI greater than .002 (Meade et al., 2008) in the CFA approach or significant chi-square difference tests in the IRT approach. This process was repeated for each of the items in the scale to identify nonequivalence and effect sizes were calculated to explore the magnitude of the effects.

[6] Note that both the IRT and CFA approaches set the parameters of the referent item to be equivalent across groups. However, in the CFA approach, the parameters are constrained to a single value (i.e., 0 and 1 for the intercept and factor loading, respectively). In contrast, the parameters for the referent item in the IRT models are freely estimated in one group and constrained to the same values in the second group.

The approaches to testing for nonequivalence were comparable but the methods of identifying nonequivalence were not. Table 1 provides a summary of the differences between the IRT and CFA approaches to examining nonequivalence. For example, the IRT approach relies solely on chi-square difference tests which will be substantially affected by sample size. In contrast, although the CFA approach has traditionally relied on chi-square differences to identify DIF, several fit indices are available for evaluating changes in fit between constrained and unconstrained models. In fact, recent research recommends using these other indices instead of chi-square differences to identify nonequivalence (Cheung and Rensvold, 2002; Meade et al., 2008). Unfortunately, comparable indices are not available in IRT and, therefore, we expect different findings regardless of the general approach to testing for DIF.

## Results

Table 2 shows means and standard deviations for each of the Big Five traits and for each decade between the ages of 20 and 50. Consistent with past research, the biggest differences across these age groups were observed on the Conscientiousness and Neuroticism scales and between the ages of 20 and 50. As described above, we examined the measurement equivalence of personality traits between the samples of 20- and 50-year olds because these comparisons provide the largest differences and, therefore, the strongest test of the artifact hypothesis.

### *IRT DIF Analyses*

Before fitting the IRT models, we first tested the IRT assumption of unidimensionality using CFA[7]. For each of the Big Five traits, we estimated three alternative models for

[7] Some authors have suggested that a linear factor model may be inappropriate for analyzing the dichotomous or polytomous (multi-categorical) data that are typically used for IRT analyses (e.g., Embretson & Reise, 2000). The issue is that categorical data violate the assumptions made by the linear factor model that the data are continuous and normally distributed. However, there are several caveats to this argument. First, these limitations primarily hold for so-called normal-theory estimators like maximum likelihood or generalized least squares. These estimators can be affected by violating the assumptions described above but other estimators, like weighted least squares (WLS) or diagonally weighted least squares (sometimes called Robust WLS), are available that specifically address this issue.

comparison. First we estimated a one-factor model to test the hypothesized structure of these unidimensional scales. Next, we also tested two alternative two-factor models. The first model was based on previous findings (Nye et al., 2008) that personality scales can be represented well by estimating separate method factors for positive and negative items. Past research has also identified a facet-level structure for each of the BFI scales with two facets underlying each trait (Soto & John, 2009). Therefore, we also tested this structure. The results of these analyses are presented in Table 3.

As shown in Table 3, the one-factor models clearly did not fit any of the BFI scales. For the Extraversion, Agreeableness, Conscientiousness, and Neuroticism scales, the two-factor models with a positive and a negative factor fit the data best[8]. However, for the Openness scale, the facet structure proposed by Soto and John (2009) fit the data substantially better than the positive/negative factor structure. Thus, the IRT assumption of unidimensionality was violated in each of the BFI scales. DIF can occur when more than one latent factor influences item responses (Camilli & Shepard, 1994). Consequently, unidimensional DIF analyses on the full BFI scales could result in some items being falsely identified as nonequivalent. We next tested

---

Second, simulation research has consistently shown that even the normal theory estimators are robust to violations of the requirement for continuous normally distributed data when there are at least five categorical response options (see Finney & DiStefano, 2006 for a review). The data examined here included five Likert response options and, therefore, will not be affected by analyzing categorical data, so we used CFA with maximum likelihood estimation to verify the dimensionality of the scales prior to estimating the IRT models.

[8] For each of the CFA models examined in Table 3, the error terms for items with similar content or that used synonyms or antonyms in item wording were allowed to correlate. For example, the error terms for the Extraversion items "Is talkative" and "Tends to be quiet" were allowed to correlate because the adjectives used in these items are antonyms. This practice is common in the personality literature (Hopwood & Donnellan, 2010) and is necessary to get accurate estimates of model parameters (Cole, Ciesla, & Steiger, 2007). However, when testing the facet structure proposed by Soto and John (2009), the relationships between these items were often modeled as a separate factor rather than with correlated errors. For example, in the Neuroticism scale, the error terms for the items "Is depressed, blue," "Is emotionally stable, not easily upset," and "Can be moody" were allowed to correlate due to their similar content when estimating positive and negative factors. However, in Soto and John's (2009) factor structure, these items were indicators of the Depression factor. Therefore, the relationships among these items were modeled by the latent factor rather than by correlated errors. This was also the case for the items "Is full of energy" and "Generates a lot of enthusiasm" on the Extraversion scale which comprised the Activity dimension in Soto & John's framework. For both of these scales (i.e., Extraversion and Neuroticism), the factor structure with positive and negative factors fit better than the facet structure even without these correlated errors.

the fit of the IRT model by calculating the chi-square fit statistics for single items and item pairs. However, due to the multidimensionality in the scales, items loading on separate dimensions were analyzed separately to assess IRT model fit.

Consistent with the work by Tackett et al. (2009), we fit the SGR model to each dimension of the BFI scales using the MULTILOG computer program (Thissen, 2003). The model-fit statistics calculated in the MODFIT computer program (Stark, 2001) suggested that the model fit the data from the 20 year olds poorly. However, because of the large sample size (N = 15,726), this finding was likely due to the sensitivity of the chi-square test to large samples. When the chi-square values were adjusted to a sample size of 500, results indicated that the SGR model fit the data well. In fact, nearly all of the adjusted $\chi^2$/df ratios for item pairs were less than 2, indicating good model fit (Chernyshenko, Stark, Drasgow, & Roberts, 2007). Therefore, we used the SGR model to examine DIF in the BFI scales. Given the multidimensionality in the scale, we estimated the multidimensional extension of the SGR model as implemented in Mplus[9] (Muthén, & Muthén, 2012). This model allowed us to estimate the parameters for all of the items (i.e., both positive and negative) simultaneously and test for DIF using the LR method.

The results of these multidimensional analyses indicated that DIF was prevalent in each of the BFI scales. In fact, our preliminary analyses to identify an equivalent referent item[10] indicated that all of the items had significant DIF. In other words, there was not an equivalent referent item that could be used to estimate DIF in the remaining items. Nonequivalent referent items will result in item parameters that are scaled differently in each of the samples and these differences can either exacerbate or mask true parameter differences between the groups. Therefore, we do not present the chi-square values here because any differences that are

[9] Example syntax for the IRT analyses is provided in the online supplementary material.
[10] More details about these analyses are provided in the Appendix.

observed would be biased upwards or downwards due to nonequivalence in the referent item. Because the IRT LR method results in a chi-square test, these results were influenced by the sample size and the Type I error rates (i.e., the number of items falsely identified as nonequivalent) of these analyses are likely high. Using samples that were substantially smaller than those examined in the present research, several studies have demonstrated this effect (Budgell, Raju, & Quartetti, 1995; Stark et al., 2006). For example, Stark et al. (2006) showed that the number of Type I errors when N = 1,000 was almost double the error rate observed in samples of 500 under some conditions. Therefore, we also examined DIF in random samples of 500 respondents from both groups (total N = 1,000). This sample size was selected because it is consistent with the sample sizes examined in previous simulation research and by Tackett et al. (2009).

Results of the IRT DIF analyses in the samples of 500 are presented in Table 4. As shown in this table, very few of the items were found to function differently across age groups relative to the results in the larger sample. The fewest items were identified in the Extraversion, Agreeableness, and Conscientiousness scales. In each of these scales, no more than two items were identified as nonequivalent. In contrast, nearly half of the items in the Neuroticism and Openness scales displayed DIF, suggesting that there were greater differences in the response processes across age groups on each of these scales.

One of the problems with using the LR method to detect DIF is that it does not demonstrate the practical importance of DIF findings. In other words, the LR method can detect even negligible differences across groups but provides very little information about the magnitude of these effects and whether or not they are substantial enough to influence the conclusions drawn from a study (Budgell, Raju, & Quartetti, 1995; Stark et al., 2006). Therefore,

we next estimated effect sizes to determine the practical importance of the DIF we identified above.

It is important to note that the effect sizes proposed by Stark et al. (2004) were developed for unidimensional scales and can only be calculated for a single dimension. Therefore, we calculated effect sizes separately for the two factors in the multidimensional BFI scales. Each item loads on only a single factor so the parameters from the multidimensional model can be used to calculate effect sizes for single items and/or dimensions. A similar approach has been used in past research to calculate effect sizes for multidimensional personality scales (Nye & Drasgow, 2011a). Using this approach, we calculated effect sizes for the unidimensional scales (c.f., Stark et al., 2004). Despite significant DIF, these differences appear to have little influence at the test level. Using the SGRDTFR computer program (Seybert, 2013), we calculated the effect size index for DTF proposed by Stark et al. (2004). Results are presented in Table 5 and suggest that the magnitude of DTF was negligible for each factor and in each of the scales. Here, positive effect sizes indicate that scores in the sample of 50 year-olds will be lower due to nonequivalence in the measure. In contrast, negative effect sizes indicate that scores in the older sample are inflated due to bias. As shown in Table 5, the largest absolute value of these effect sizes was -.16 for the negative factor in the Agreeableness scale. Using Cohen's (1988) guidelines, this effect size would be considered negligible which suggests that differential functioning has very little influence on score comparisons across these two age groups.

Remember also that the observed differences between groups can be broken down into the effects of bias and the effects of impact. Consequently, once the effects of DTF have been determined, we can calculate the true differences in the latent trait over time. Table 5 also provides the effect sizes for the observed differences and for impact between age groups. As

shown in this table, the overall differences between groups were largest for conscientiousness, neuroticism, and the negative dimension of agreeableness. In fact, we found substantial observed differences between age groups on the positive (d = -.48) and negative (d = -.62) conscientiousness factors. In addition, these differences were almost entirely due to impact (true differences between groups) and the same was true for the positive dimension of neuroticism. For the negative dimensions of agreeableness and neuroticism, DTF accounted for a larger portion of the observed difference than impact. However, the effect size of DTF in both cases was still negligible, suggesting that despite the larger relative effect, DTF still had very little influence on observed differences.

*MACS Analyses*

We next tested the equivalence of the BFI scales using MACS analyses conducted with maximum likelihood estimation[11] in Mplus 7[12]. Again, the goal of these analyses was to replicate the results provided above using an alternative method and to illustrate the similarities and differences between these two approaches to testing for measurement equivalence. Therefore, we re-analyzed the same samples of 15,726 and 500 individuals using MACS analyses.

For each of the Big Five traits, the configural model fit the data well when the best fitting models from Table 3 were estimated simultaneously in each group, suggesting that the same factor structure was appropriate for both samples. Therefore, consistent with current

[11] Bollen (1989) and others (e.g., Kaplan, 2000) have noted that Likert-type data provides ordered-categorical responses that violate the assumption of normality in maximum likelihood estimation. As such, ordinal estimation methods like weighted least squares (WLS) or diagonally weighted least squares (DWLS) could be used to model Likert responses. However, this approach may be less useful when testing for measurement equivalence. Nye and Drasgow (2011b) showed that some of the fit indices calculated with ordinal estimation methods were unable to detect misfit. In fact, the CFI used to identify nonequivalence was particularly insensitive to model misfit and rarely varied from 1.00 even when the model was clearly misspecified. Consequently, ordinal estimation methods may not be useful for detecting DIF. Fortunately, simulation research has shown that using maximum likelihood estimation to model Likert responses with at least five categories has minimal effect on fit indices, parameter estimates, and standard errors (Bollen, 1989; Finney & DiStefano, 2006; Muthén & Kaplan, 1985). Therefore, we chose to use maximum likelihood estimation instead of the alternative ordinal estimation methods.

[12] Example syntax for the CFA analyses is provided in the online supplementary material.

recommendations for testing measurement nonequivalence (cf. Stark et al., 2006), we next tested metric and scalar equivalence simultaneously for each item. The results of the DIF analyses using CFA in the sample of 15,726 provided a very different picture of DIF than the IRT analyses in the same sample. The IRT analyses suggested that all of the items had significant DIF but very few of the items displayed DIF across age groups using the CFA approach. Using the ΔCFI criterion to identify DIF, only one item in the Agreeableness scale and two items in the Conscientiousness scale showed DIF across age groups. The most DIF was identified in the Neuroticism and Openness scales where three and four items, respectively, in each of these scales showed DIF. Overall, the majority of the items were equivalent across groups.

Given that DIF was identified in some of the items, we also examined the effect size of this DIF. The item-level effect sizes were generally below .20, suggesting only small effects. The largest effect size was .26 for the item "Has an assertive personality" in the Extraversion scale. A few Openness items also had effect sizes around .22 and .23. However, the overall magnitude of these effects was still small (Cohen, 1988). Full results for these analyses including both effect sizes and the fit indices for the constrained models are provided in the Appendix.

Using the equations proposed by Nye and Drasgow (2011a), we also calculated the effect sizes of measurement nonequivalence on scale-level means. The effect sizes of the observed differences, the differences due to nonequivalence, and the differences due to impact are provided in Table 6. Not surprisingly, the results showed that the overall effects of nonequivalence on the means were also small. Here, the largest effects were observed for the positively worded items in the Extraversion ($d$ = -.13) and Conscientiousness ($d$ = .12) scales and the negatively worded items in the Agreeableness scale ($d$ = -.13). However, these effect sizes would still be considered negligible effects. In fact, in some cases, measurement nonequivalence

masked part of the true differences between groups. For example, although the effect size of the observed differences for the positively worded Extraversion items was only .06, we found that the actual effect size was around .19 after correcting for the influence of DTF. Similarly, the effect size for true differences (i.e., impact) on the positively worded Conscientiousness items was actually larger than the observed differences due to the effects of DTF. This masking effect occurs because the sign of the observed difference is in the opposite direction of the effect of DIF. Specifically, although the observed differences on the positively worded Conscientiousness items indicated higher scores in the older sample, nonequivalence inflated scores in the younger sample which created smaller observed differences between groups. Consequently, after removing the effects of nonequivalence, the size of some differences increased.

For comparison with the IRT results, we also conducted MACS analyses on the sample of 500 respondents. The results are provided along with the IRT results for this sample in Table 4. As shown, the results were largely consistent with the results in the larger sample. In addition, the effect sizes were generally small, suggesting that DIF had only negligible effects on observed differences between groups. There were some differences in the number of DIF items identified in the larger sample and the reduced sample but these differences were likely due to sampling variation and the small overall effects of DIF. Table 7 also provides the effects sizes of observed differences, differences due to DTF, and true differences (i.e., impact). The results in Table 7 indicate that the majority of the observed differences between these two age groups can be attributed to impact rather than DTF. Again, these results are consistent with the CFA results in the larger sample.

## Discussion

Recent research suggests that age differences in personality traits could be due to measurement nonequivalence across ages rather than true differences in the latent traits. Using large samples and comparing responses across ages 20 and 50, our results suggest that this is not the case. A limited amount of DIF was identified using IRT and CFA methodologies but these differences had very small effects on mean-level comparisons across age groups. The largest effect size (in absolute value) for DTF was only -.16 in the IRT analyses and the majority of the effect sizes were below .10. Even after the effects of DIF were removed, mean level comparisons suggested that older individuals were more conscientious and less neurotic than their younger counterparts. These differences are consistent with the age differences found in past research. Results also indicated that the older sample was more agreeable and open to experiences, but these effects were somewhat smaller.

Importantly, the effect sizes of the true differences between age groups on the Conscientiousness scale were substantial. It is possible that both individual change and cohort effects are contributing to these differences. Past research has provided evidence for both sources of age differences in personality traits but has also indicated that neither personality development nor cohort effects alone completely explain the differences that are observed across age groups in cross-sectional research (Roberts et al., 2006; Smits et al., 2011). As a result, both factors may have contributed to these differences and enhanced the magnitude of the effects across ages.

In the present study, different results were obtained when using IRT and CFA analyses in the large sample. Although all of the items were identified as nonequivalent using the IRT LR approach, the majority of items were found to be equivalent using CFA MACS analyses. The primary reason for these differences was that the IRT approach is limited to the use of chi-square difference tests which are extremely sensitive to sample size. Consequently, the results of the

IRT analyses reflect both the magnitude of the differences between groups and the characteristics of the sample (i.e., the sample size). This focus on chi-square tests limits the conclusions that can be drawn from these analyses and the generalizability of the findings. In contrast, a number of fit indices have been developed for CFA and are available for testing nonequivalence. These indices were developed to be less affected by sample size and past research has generally confirmed that this is the case (Cheung & Rensvold, 2002; Meade et al., 2008; Nye, 2011). As a result, evaluating nonequivalence with the CFA methodology may provide more accurate results in larger samples. However, it is worth noting that the results from the CFA analyses would have been more consistent with the IRT results if the chi-square difference test had been used[13]. Therefore, these results provide an illustration of the benefit of using alternative indicators of nonequivalence.

One possible way of addressing the sample size issue with IRT analyses is by calculating effect sizes to help facilitate the interpretation of DIF results. Although the chi-square difference tests identified several items as nonequivalent even in the sample of 500, the effect sizes suggested that the effects were small and had little effect on observed differences. Similar results were obtained with the CFA analyses. As such, effect sizes for both techniques can provide additional information about DIF or DTF that may be identified in a scale. However, accurate estimates of the effect size can be calculated only if an equivalent referent item can be identified to set the metric of the latent factor and the estimated parameters. This was not the case for the IRT analyses in the sample of 15,726 and, therefore, effect sizes could not be estimated for the IRT analyses in the larger sample. Another way of addressing the sensitivity of the IRT analyses

[13] It is also important to note that the accuracy of DIF detection when using the chi-square difference test in both the CFA and IRT approaches is dependent on the fit of the baseline model. When the baseline model for IRT (i.e., the unconstrained model) or CFA (i.e., the configural model) does not adequately fit the data, comparisons of constrained and unconstrained models may provide inaccurate results (e.g., Maydeu-Olivares & Cai, 2006; Yuan & Bentler, 2004).

to sample size is to develop alternative fit indices that are analogous to those developed for CFA. Although alternative indices are starting to be developed (e.g., Maydeu-Olivares & Joe, 2014; Maydeu-Olivares & Liu, 2014), these indices are not yet widely available in many commonly used statistical packages and, in some cases, are influenced by the characteristics of the data. Therefore, more research is needed to evaluate the accuracy of these indices and to understand their interpretation. In addition, many of the fit indices used for evaluating CFA model fit have been found to be poor indicators of DIF when compared across nested models (Cheung & Renvold, 2002; Meade et al., 2008). Consequently, simulation research is also needed to evaluate the accuracy of the new IRT fit indices for identifying DIF.

*Limitations*

This study was limited by the use of cross-sectional data. Given the inherent difficulty with following a large group of individuals over a 30-year period, collecting longitudinal data to examine differences across these age groups would be difficult. Therefore, we were limited to cross-sectional comparisons across young and middle-aged adults and these data affect the conclusions that we can draw about individual-level change and the effects of time on self-reports of personality traits. Moreover, a growing body of research suggests that there can be substantial intra-individual variability in both the level and structure of personality over time (Nesselroade, Gerstorf, Hardy, & Ram, 2007; Nesselroade & Molenaar, 2010) and this level of variability cannot be addressed with the current data either. However, the goal of this study was to address concerns about measurement nonequivalence across age-groups, for which cross-sectional data were appropriate. As such, these results provide evidence that DIF has only a negligible effect on comparisons of BFI scores across age groups.

Another limitation of these cross-sectional data is that our analyses cannot determine whether the differences that we identified are due to developmental trends or cohort effects. In the literature on personality trait change, there is empirical evidence for both within-person change and cohort differences (e.g., Roberts, Edmonds, & Grijalva, 2010; Roberts et al., 2006; Smits et al., 2011; Terracciano, McCrae, Brant, & Costa, 2005). Comparisons of age groups in the present study are potentially influenced by both as well. A previous study conducted by Soto et al. (2011) did examine cohort differences in the same sample used in the present study and found little evidence of a pronounced cohort effect. However, their analyses focused on a much shorter interval between age groups (i.e., 7 years) than the 30-year age differences we examine here. Moreover, they could not differentiate cohort effects from within-person change because the data are not longitudinal.

It is worth noting that the identification of both developmental trends and cohort effects would be affected by measurement nonequivalence. In both cases, comparisons are being made across groups or over time and, therefore, measurement nonequivalence could potentially influence mean-level results. Consequently, the present study demonstrates that comparisons between 20- and 50-year olds are not substantially affected by measurement nonequivalence, regardless of whether these differences are due to developmental trends or cohort differences. Therefore, more research is needed to differentiate these effects and to explain the potential sources of differences over time or across age groups.

Finally, another limitation of this study is that we were not able to determine the source of the limited amount of DIF that we did identify. Although several items were identified as nonequivalent, their effects had little impact on group-level mean comparisons. Looking at these items, it is not immediately clear why the responses to these items varied between age groups

and there are a number of potential reasons for the differences we identified. DIF could be due to age differences in item interpretation, the frame-of-reference used when responding to items, life experiences, or any number of alternative potential causes. However, it is important to note that measurement nonequivalence is not always a negative characteristic in a study. Nonequivalence is frequently termed an "artifact" but differences in the response process may reflect the same substantive changes that are the source of age differences in personality traits (Nye & Roberts, 2013). For example, biological processes related to aging or increased investments in an individual's work, family, or social roles have been hypothesized to be contributing factors to personality trait change (Bleidorn, Klimstra, Denissen, Rentfrow, Potter, & Gosling, 2013; Lodi-Smith & Roberts, 2007; McCrae & Costa, 2003) and may result in measurement nonequivalence across age groups. In other words, age differences in personality traits can take many forms and, therefore, identifying measurement nonequivalence may not be as important as understanding why these differences have been observed. As a result, future research should attempt to understand the sources of nonequivalence and to identify whether or not these differences are response artifacts or meaningful individual change.

However, efforts to identify the sources of nonequivalence will be limited by the absence of a theoretical framework for understanding nonequivalence or its causes. The current methods of testing for invariance (whether in the IRT or CFA framework) can identify nonequivalence but provide little information for predicting when and if it will occur. More recent methods of testing for DIF (e.g., Wood, 2009) provide ways to identify potential causes. Still, the lack of theory in this area precludes the use of strong tests (Platt, 1964) for nonequivalence that involve developing hypotheses and testing these hypotheses in a controlled experimental or quasi-experimental research design. With this in mind, future research on the equivalence of

personality trait scores across ages or over time would greatly benefit from a taxonomy of nonequivalence and a theoretical description of when such types of nonequivalence might be observed.

## Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality*, *75*, 323-358.

Allemand, M., Zimprich, D., & Hendriks, A. A. (2008). Age differences in five personality domains across the life span. *Developmental Psychology*, *44*, 758-770.

Benet-Martínez, V., & John, O. P. (1998). Los cinco grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, *75*, 729-750.

Bleidorn, W., Kandler, C., Riemann, R., Angleitner, A., & Spinath, F. M. (2009). Patterns and sources of adult personality development: Growth curve analyses of the NEO PI-R scales in a longitudinal twin study. *Journal of Personality and Social Psychology, 97,* 142–155.

Bleidorn, W., Klimstra, T. A., Denissen, J. J. A., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality maturation around the world—A cross-cultural examination of Social Investment Theory. *Psychological Science*, *24*, 2530-2540.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement*, *19*, 309−324.

Byrne, B. M. (1998). *Structural equation modeling in LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahweh, NJ: Lawrence Erlbaum Associates.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*, 88-106.

Cheung, G. W., & Rensvold, R. B. (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233−255.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, *12*, 381-398.

Costa, P. T., & McCrae, R. R. (1988). Personality in adulthood: a six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology*, *54*, 853-863.

Costa, P. T. Jr., & McCrae, R. R. (1992). *Professional manual: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI).* Odessa, FL: Psychological Assessment Resources.

Donnellan, M. B., Conger, R. D., & Burzette, R. G. (2007). Personality development from late adolescence to young adulthood: Differential stability, normative maturity, and evidence for the maturity-stability hypothesis. *Journal of Personality*, *75*, 237-264.

Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: evidence from two national samples. *Psychology and Aging*, *23*, 558-566.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*, 134−135.

Drasgow, F., Levine M. V., Tsien, S., Williams B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. D. Mueller (Eds.), *Structural equation modeling: A second course.* (pp. 269-314). Greenwich, CT: Information Age Publishing.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275-299.

Gonzalez-Roma, V., Hernandez, A., & Gomez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential items functioning in graded response items. *Multivariate Behavioral Research*, *41*, 29-53.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93-104.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated?. *Personality and Social Psychology Review*, *14*, 332-346.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *3*, 424−453.

Jackson, J. J., Bogg, T., Walton, K., Wood, D., Harms, P. D., Lodi-Smith, J. L., & Roberts, B. W. (2009). Not all conscientiousness scales change alike: A multi-method, multi-sample study of age differences in the facets of conscientiousness. *Journal of Personality and Social Psychology, 96,* 446-459.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big Five Inventory: Technical report. *Berkeley: University of California, Berkeley.*

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*: Thousand Oaks, CA: Sage.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford Press.

Labouvie-Vief, G., Diehl, M., Tarnowski, A., & Shen, J. (2000). Age differences in adult personality: Findings from the United States and China. *Journal of Gerontology Series B*, *55*, 4-17.

Lei, M., & Lomax, R. G. (2005). The effects of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*, 1−27.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53- 76.

Lodi-Smith, J.L. & Roberts, B.W. (2007). Social investment and personality: A meta-analytic analysis of the relationship of personality traits to investment in work, family, religion, and volunteerism. *Personality and Social Psychology Review*, *11*, 68-86.

Lodi-Smith, J., & Roberts, B. W. (2010). Getting to know me: Social role experiences and age differences in self-concept clarity during adulthood. *Journal of Personality*, *78*, 1383-1410.

Lucas, R. E., & Donnellan, M. B. (2009). Age differences in personality: Evidence from a nationally representative Australian sample. *Developmental Psychology*, *45*, 1353-1363.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*, *40*, 261-279.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55-64.

Maydeu-Olivares, A., & Liu, Y. (2014). Item diagnostics in multivariate discrete data. Manuscript in preparation.

McCrae, R. R., & Costa, P. T., Jr. (2003). *Personality in adulthood: A Five-Factor theory perspective*. New York, NY: Guilford Press.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568−592.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. British *Journal of Mathematical and Statistical Psychology*, *38*, 171-189.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Focus article: Idiographic filters for psychological constructs. *Measurement*, *5*, 217-235.

Nesselroade, J. R., & Molenaar, P. C. (2010). Analyzing intra-person variation: Hybridizing the ACE model with P-technique factor analysis and the idiographic filter. *Behavior genetics*, *40*, 776-783.

Nye, C. D., (2011). *The development and validation of effect size measures for IRT and CFA studies of measurement equivalence.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Nye, C. D., & Drasgow, F. (2011a). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*, 966-980.

Nye, C. D., & Drasgow, F. (2011b). Assessing goodness of fit: Simple rules of thumb simply don't work. *Organizational Research Methods*, *14*, 548−570.

Nye, C. D., & Roberts, B. W. (2013). A developmental perspective on the importance of personality for understanding workplace behavior. In N. Christiansen & R. Tett (Eds.), *Handbook of Personality at Work* (pp. 796-818). New York: Routledge.

Nye, C., Roberts, B. W. Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality, 42,* 1524-1536.

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347-353.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517−529.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Rentfrow, P. J., Gosling, S. D., & Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science*, *3*, 339-369.

Roberts, B. W., Edmonds, G., & Grijalva, E. (2010). It is developmental me, not Generation Me developmental changes are more important than generational changes in Narcissism—Commentary on Trzesniewski & Donnellan (2010). *Perspectives on Psychological Science*, *5*, 97-102.

Roberts, B. W., Walton, K. & Viechtbauer, W. (2006). Patterns of mean-Level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1-25.

Roberts, B. W., & Wood, D. (2006). Personality development in the context of the Neo-Socioanalytic Model of personality. In D. Mroczek & T. Little (Eds.), *Handbook of Personality Development* (pp. 11-39). Mahwah, NJ: Lawrence Erlbaum Associates.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometric Monograph No. 18). Iowa City, IA: Psychometric Society.

Satorra, A., & Bentler, P. M. (1999). *A scaled difference chi-square test statistic for moment structure analysis*. Retrived February 17, 2015 from University of California, Los Angeles (UCLA), Department of Statistics Web site: http://statistics.ucla.edu/preprints/uclastat-preprint-1999:19.

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63*, 506-516.

Seybert, J. (2013). *SGRDTFR* [computer program]. Department of Psychology, University of South Florida.

Small, B. J., Hertzog C., Hultsch D. F., & Dixon R. A. (2003). Stability and change in adult personality over 6 years: Findings from the Victoria Longitudinal Study. *Journal of Gerontology: Psychological Sciences*, 58, 166–176.

Smits, I. A. M., Dolan, C. V., Vorst, H. C. M., Wicherts, J. M., & Timmerman, M. E. (2011). Cohort differences in Big Five personality factors over a period of 25 years. *Journal of Personality and Social Psychology*, *100*, 1124-1138.

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84-90.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*, 330-348.

Specht, J., Egloff, B., & Schmukle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, *101*, 862-882.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041-1053.

Stark, S. (2001). *MODFIT: A computer program for model-data fit* [computer program]. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*, 497−508.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292−1306.

Tackett, J. L., Balsis, S., Oltmanns, T. F., & Krueger, R. F. (2009). A unifying perspective on personality pathology across the life span: Developmental considerations for the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders. *Development and Psychopathology*, *21*, 687-713.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T., Jr. (2005). Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging*, *20*, 493-506.

Thissen, D. (2003). *MULTILOG* 7: *Multiple categorical item analysis and test scoring using item response theory* [computer program]. Chicago, IL: Scientific Software.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147−169). Hillsdale, NJ: Erlbaum.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1-27.

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737-757.

Table 1

*Summary of the Differences between the IRT and CFA Approaches to Identifying DIF*

| | **Item Response Theory (IRT)** | **Confirmatory Factor Analyses (CFA)** |
|---|---|---|
| Relationship between the latent trait and item responses | Assumed to be non-linear | Assumed to be linear |
| Baseline model comparisons | Tests whether the same IRT model fits in both groups | Tests whether the same factor structure fits in both groups |
| Parameters constrained to test for DIF | $\boldsymbol{a_i}$ (item discrimination parameter)<br>$\boldsymbol{b_{ik}}$ (item difficulty parameter) | $\boldsymbol{\lambda_i}$ (item factor loading)<br>$\boldsymbol{\tau_i}$ (item intercept) |
| Model fit indices | $\chi^2$ for single items and item pairs | Several indices are available: $\chi^2$, RMSEA, NNFI/TLI, CFI, SRMR |
| Criteria for identifying DIF | A significant likelihood ratio test (i.e., $\chi^2$ difference test) indicates DIF | $\Delta$CFI > .002 indicates nonequivalence |
| Effect sizes for DIF | $d_{DTF}$ (Stark, Chernyshenko, & Drasgow, 2004) | $d_{MACS}$ (Nye & Drasgow, 2011) |
| Key differences for DIF research | • The concept of DTF and the compensatory nature of DIF is more clearly addressed with IRT<br>• The logistic model underlying IRT is considered more appropriate for dichotomous items (Raju, Laffitte, & Byrne, 2002) | • Multiple fit indices are available for examining model fit and identifying DIF<br>• Multiple latent traits can be examined more efficiently<br>• Stricter forms of invariance (e.g., equivalent error variances or relationships with other variables) can be examined within this framework |

Table 2

*Descriptive Statistics for the BFI Scales in Each Sample*

| Item | Age 20 Mean (SD) | Age 30 Mean (SD) | Age 40 Mean (SD) | Age 50 Mean (SD) |
|---|---|---|---|---|
| Extraversion | 3.25 (.81) | 3.25 (.82) | 3.27 (.82) | 3.25 (.81) |
| Agreeableness | 3.63 (.65) | 3.62 (.65) | 3.71 (.65) | 3.80 (.64) |
| Conscientiousness | 3.39 (.69) | 3.57 (.69) | 3.70 (.69) | 3.78 (.67) |
| Neuroticism | 3.05 (.83) | 2.99 (.83) | 2.90 (.83) | 2.83 (.82) |
| Openness | 3.71 (.63) | 3.80 (.64) | 3.77 (.65) | 3.82 (.67) |

Note: Age 20 N = 142,605; Age 30 N = 54,555; Age 40 N = 26,803; Age 50 N = 15,726.

Table 3

*Results from the CFA Analyses of Alternative Models for Each Scale*

| Big 5 Factor | RMSEA | CFI | NNFI | SRMR |
|---|---|---|---|---|
| **Agreeableness** | | | | |
| 1-Factor | .08 | .89 | .85 | .04 |
| **2-Factor (pos./neg.)** | **.05** | **.95** | **.94** | **.03** |
| 2-Factor (facets) | .08 | .89 | .85 | .04 |
| **Conscientiousness** | | | | |
| 1-Factor | .10 | .89 | .86 | .05 |
| **2-Factor (pos./neg.)** | **.05** | **.98** | **.97** | **.02** |
| 2-Factor (facets) | .10 | .89 | .85 | .05 |
| **Extraversion** | | | | |
| 1-Factor | .10 | .94 | .90 | .04 |
| **2-Factor (pos./neg.)** | **.04** | **.99** | **.99** | **.02** |
| 2-Factor (facets) | .10 | .94 | .90 | .04 |
| **Neuroticism** | | | | |
| 1-Factor | .10 | .92 | .87 | .04 |
| **2-Factor (pos./neg.)** | **.06** | **.98** | **.96** | **.03** |
| 2-Factor (facets) | .10 | .92 | .88 | .04 |
| **Openness** | | | | |
| 1-Factor | .08 | .89 | .86 | .05 |
| 2-Factor (pos./neg.) | .08 | .89 | .85 | .05 |
| **2-Factor (facets)** | **.04** | **.98** | **.97** | **.03** |

Note: Because the goal of our paper was to test for nonequivalence across age groups and nonequivalence can be item-specific, we chose to include all of the items in the BFI scale rather than excluding items to estimate each of the facets as suggested by Soto and John (2009). Excluding these items would have limited the conclusions that we could draw about the measurement of these constructs over time. Therefore, we classified excluded items into the facets suggested by Soto and John based on content and estimated the factor structure using the full scale.

Table 4

*IRT and CFA DIF Results for the BFI Scales in the Sample of 500*

| | IRT Analyses[b] | CFA Analyses | | | | |
|---|---|---|---|---|---|---|
| **Big 5 Factor** | **$\Delta\chi^2$** | **$\chi^2$ (df)** | **RMSEA** | **CFI** | **NNFI** | **$d_{MACS}$** |
| **Extraversion** | | | | | | |
| 2-Factor (Age 20) | | 16.19 (17) | .00 | 1.00 | 1.00 | |
| 2-Factor (Age 50) | | 34.88 (17) | .05 | .99 | .99 | |
| **Configural Invariance** | | 51.63 (34) | .03 | .994 | .99 | |
| **Scalar Invariance** | | | | | | |
| ...Is talkative | 7.87 | 52.25 (36) | .03 | .994 | .99 | .07 |
| ...Is full of energy[a] | -- | -- | -- | -- | -- | -- |
| ...Generates a lot of enthusiasm | 10.84 | 57.14 (36) | .03 | .992 | .99 | .15 |
| ...Has an assertive personality | 13.36* | 59.49* (36) | .04 | .992 | .99 | .22 |
| ...Is outgoing, sociable | 9.72 | 53.32 (36) | .03 | .994 | .99 | .12 |
| ...Is reserved[a] | -- | -- | -- | -- | -- | -- |
| ...Tends to be quiet | 7.51 | 54.32 (36) | .03 | .993 | .99 | .12 |
| ...Is sometimes shy, inhibited | 15.04* | 55.83 (36) | .03 | .993 | .99 | .16 |
| **Agreeableness** | | | | | | |
| 2-Factor (Age 20) | | 30.90 (25) | .02 | 1.00 | .99 | |
| 2-Factor (Age 50) | | 41.38 (25) | .03 | .99 | .98 | |
| **Configural Invariance** | | 125.76 (52) | .05 | .954 | .94 | |
| **Scalar Invariance** | | | | | | |
| ...Tends to find fault with others | 2.01 | 129.08 (54) | .05 | .953 | .94 | .17 |
| ...Is helpful and unselfish with others | 2.45 | 129.12 (54) | .05 | .953 | .94 | .12 |
| ...Starts quarrels with others | 9.79 | 131.67 (54) | .05 | .951** | .94 | .24 |
| ...Has a forgiving nature | 4.02 | 126.91 (54) | .05 | .954 | .94 | .07 |
| ...Is generally trusting | 9.01 | 139.88* (54) | .06 | .946** | .93 | .26 |
| ...Can be cold and aloof[a] | -- | -- | -- | -- | -- | -- |
| ...Is considerate and kind to almost everyone[a] | -- | -- | -- | -- | -- | -- |
| ...Is sometimes rude to others | 8.03 | 131.20 (54) | .05 | .951** | .94 | .25 |
| ...Likes to cooperate with others | 12.92* | 129.47 (54) | .05 | .952 | .94 | .14 |
| **Conscientiousness** | | | | | | |
| 2-Factor (Age 20) | | 63.73 (26) | .06 | .98 | .97 | |
| 2-Factor (Age 50) | | 54.23 (26) | .04 | .99 | .98 | |
| **Configural Invariance** | | 118.19 (52) | .05 | .968 | .96 | |
| **Scalar Invariance** | | | | | | |
| ...Does a thorough job | 2.79 | 121.27 (54) | .05 | .968 | .96 | .16 |
| ...Can be somewhat careless[a] | -- | -- | -- | -- | -- | -- |
| ...Is a reliable worker | 10.14 | 123.47 (54) | .05 | .967 | .96 | .19 |
| ...Tends to be disorganized | 19.03* | 131.07* (54) | .05 | .963** | .95 | .24 |
| ...Tends to be lazy | 13.81* | 119.18 (54) | .05 | .969 | .96 | .10 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ...Perseveres until the task is finished[a] | -- | -- | -- | -- | -- | -- |
| ...Does things efficiently | 8.54 | 122.94 (54) | .05 | .967 | .96 | .16 |
| ...Makes plans and follows through with them | 8.26 | 119.63 (54) | .05 | .969 | .96 | .12 |
| ...Is easily distracted | 2.85 | 119.87 (54) | .05 | .969 | .96 | .14 |
| **Neuroticism** | | | | | | |
| 2-Factor (Age 20) | | 39.88 (16) | .055 | .98 | .97 | |
| 2-Factor (Age 50) | | 42.41 (16) | .057 | .98 | .97 | |
| **Configural Invariance** | | 82.28 (32) | .06 | .980 | .97 | |
| **Scalar Invariance** | | | | | | |
| ...Is depressed, blue[a] | -- | -- | -- | -- | -- | -- |
| ...Can be tense | 1.01 | 82.53 (34) | .05 | .981 | .97 | .05 |
| ...Worries a lot | 10.71 | 83.56 (34) | .05 | .981 | .97 | .10 |
| ...Can be moody | 14.45* | 93.95* (34) | .06 | .977** | .96 | .25 |
| ...Gets nervous easily | 12.50* | 91.71* (34) | .06 | .977** | .96 | .25 |
| ...Is relaxed, handles stress well | 9.42 | 90.77* (34) | .06 | .978 | .96 | .25 |
| ...Is emotionally stable, not easily upset | 12.84* | 87.10 (34) | .06 | .979 | .96 | .15 |
| ...Remains calm in tense situations[a] | -- | -- | -- | -- | -- | -- |
| **Openness** | | | | | | |
| 2-Factor (Age 20) | | 61.10 (33) | .041 | .985 | .98 | |
| 2-Factor (Age 50) | | 55.68 (33) | .036 | .990 | .99 | |
| **Configural Invariance** | | 118.10 (66) | .04 | .978 | .97 | |
| **Scalar Invariance** | | | | | | |
| ...Is original comes up with new ideas[a] | -- | -- | -- | -- | -- | -- |
| ...Is curious about many different things | 10.38 | 118.68 (68) | .04 | .978 | .97 | .05 |
| ...Is ingenious, a deep thinker | 1.96 | 121.44 (68) | .04 | .977 | .97 | .17 |
| ...Has an active imagination | 7.41 | 123.70 (68) | .04 | .976 | .97 | .17 |
| ...Is inventive | 12.87* | 123.89 (68) | .04 | .976 | .97 | .14 |
| ...Prefers work that is routine | 17.81* | 124.70* (68) | .04 | .976 | .97 | .21 |
| ...Likes to reflect, play with ideas | 7.34 | 119.15 (68) | .04 | .978 | .97 | .10 |
| ...Values artistic, aesthetic experiences[a] | -- | -- | -- | -- | -- | -- |
| ...Has few artistic interests | 18.40* | 127.73* (68) | .04 | .974** | .97 | .24 |
| ...Is sophisticated in art, music, or literature | 15.22* | 127.03* (68) | .04 | .975** | .97 | .22 |

Notes: *p < .05, **ΔCFI > .002.

[a] Referent items.

[b] Chi-square difference tests cannot be calculated directly for the IRT analyses using the model based chi-squares reported by Mplus (Satorra & Bentler, 1999). Therefore, Satorra and Bentler (1999) derived equations for calculating a chi-square difference test using the loglikelihood of the model and these equations were used to conduct significance tests for the IRT DIF analyses. Because the model based chi-square statistics were not used directly, we do not report them here and instead only report the chi-square differences for the IRT analyses.

Table 5

*Effect Sizes for the Observed Differences, Differences Due to IRT DTF, and Impact for Each BFI Scale in the Sample of 500*

| | Factor 1 | | | Factor 2[a] | | |
|---|---|---|---|---|---|---|
| **Item** | **Observed Differences** | **DTF** | **Impact** | **Observed Differences** | **DTF** | **Impact** |
| Extraversion | .06 | .00 | .06 | .06 | .02 | .05 |
| Agreeableness | -.14 | -.08 | -.07 | -.23 | -.16 | -.07 |
| Conscientiousness | -.48 | .00 | -.48 | -.62 | .06 | -.67 |
| Neuroticism | .21 | -.03 | .24 | .21 | .15 | .07 |
| Openness | -.19 | -.05 | -.15 | -.16 | .01 | -.17 |

Note: For Extraversion, Agreeableness, Conscientiousness, and Neuroticism, Factor 1 represents the factors for the positively worded items and Factor 2 represents the factors for the negatively worded items. For the Openness scale, Factor 1 represents the Ideas facet and Factor 2 represents the Aesthetics facet from Soto and John (2009). The sum of effect sizes for DTF and Impact do not always equal the effect for observed differences due to rounding error.

[a] Negatively-worded items were reverse scored so that higher scores reflect higher levels of the latent trait to facilitate interpretation of effect sizes across Factor 1 and Factor 2.

Table 6

*Effect Sizes for the Observed Differences, Differences Due to CFA DTF, and Impact for Each BFI Scale in the Sample of 15,726*

| | Factor 1 | | | Factor 2[a] | | |
|---|---|---|---|---|---|---|
| **Item** | **Observed Differences** | **DTF** | **Impact** | **Observed Differences** | **DTF** | **Impact** |
| Extraversion | .06 | -.13 | .19 | -.09 | .00 | -.09 |
| Agreeableness | -.15 | -.04 | -.11 | -.30 | -.13 | -.17 |
| Conscientiousness | -.44 | .12 | -.56 | -.57 | -.03 | -.55 |
| Neuroticism | .21 | -.06 | .27 | .27 | .11 | .16 |
| Openness | -.19 | -.01 | -.18 | -.07 | .07 | -.13 |

Note: For Extraversion, Agreeableness, Conscientiousness, and Neuroticism, Factor 1 represents the factors for the positively worded items and Factor 2 represents the factors for the negatively worded items. For the Openness scale, Factor 1 represents the Ideas facet and Factor 2 represents the Aesthetics facet from Soto and John (2009). The sum of effect sizes for DTF and Impact do not always equal the effect for observed differences due to rounding error.
[a] Negatively-worded items were reverse scored so that higher scores reflect higher levels of the latent trait to facilitate interpretation of effect sizes across Factor 1 and Factor 2.

Table 7

*Effect Sizes for the Observed Differences, Differences Due to CFA DTF, and Impact for Each BFI Scale in the Sample of 500*

| | Factor 1 | | | Factor 2[a] | | |
|---|---|---|---|---|---|---|
| **Item** | **Observed Differences** | **DTF** | **Impact** | **Observed Differences** | **DTF** | **Impact** |
| Extraversion | .06 | -.04 | .10 | .06 | .01 | .05 |
| Agreeableness | -.14 | .01 | -.15 | -.23 | -.10 | -.13 |
| Conscientiousness | -.48 | .06 | -.54 | -.62 | .03 | -.65 |
| Neuroticism | .27 | .13 | .14 | .20 | -.01 | .22 |
| Openness | -.19 | .01 | -.20 | -.16 | .03 | -.19 |

Note: For Extraversion, Agreeableness, Conscientiousness, and Neuroticism, Factor 1 represents the factors for the positively worded items and Factor 2 represents the factors for the negatively worded items. For the Openness scale, Factor 1 represents the Ideas facet and Factor 2 represents the Aesthetics facet from Soto and John (2009). The sum of effect sizes for DTF and Impact do not always equal the effect for observed differences due to rounding error.
[a] Negatively-worded items were reverse scored so that higher scores reflect higher levels of the latent trait to facilitate interpretation of effect sizes across Factor 1 and Factor 2.

# Appendix

## *IRT DIF Analyses*

As described in the manuscript, IRT describes an individual's responses to a personality assessment as a function of the characteristics of the items and his or her standing on the latent trait. This relationship can be represented graphically in an item characteristic curve (ICC) like the ones shown in Figure A1. Here, the x-axis represents a range of latent trait values (symbolized by the Greek letter theta, θ) and the y-axis represents the probability of endorsing the item.

For the SGR model used in this study, the probability that an individual will endorse a particular option in a Likert scale is defined by

$$P(x=k)=\frac{1}{1+e^{-a_i(\theta-b_{ik-1})}}-\frac{1}{1+e^{-a_i(\theta-b_{ik})}}$$

where $P(x=k)$ is the probability that response $x$ will equal category $k$, $a_i$ is the discrimination parameter which describes the utility of the item for differentiating between high and low levels of the latent trait, and $b_{ik}$ is the location (or difficulty) parameter for option $k$ on item $i$, which defines the point on the ICC where the probability of endorsing option $k$ is .50. Thus, the SGR model estimates $m-1$ $b$-parameters for each item where $m$ is the total number of response options. Therefore, a measure with a five-point response scale will have one $a$-parameter and four $b$-parameters. As a result, the SGR model defines an ICC like the one illustrated in Figure A2 and each line represents the probability of endorsing a particular response option at each level of the latent trait.

## *IRT Model Fit*

Chernyshenko, Stark, Drasgow, and Roberts (2007) suggested two primary aspects of IRT model fit that must be addressed. First, the assumptions of the model must be consistent

with the dimensionality of the data. Many IRT models assume that the data are unidimensional but multidimensional IRT models are also available. Therefore, the dimensionality of the scale must be identified before IRT models can be applied. In fact, Maydeau-Olivares (2005) suggested that issues with multidimensionality may have been the source of misfit in previous research using IRT with personality scales.

After verifying the dimensionality of the data, fit should also be assessed by comparing the predictions made by the model with observed responses. If the IRT model adequately describes the response process, the expected number of individuals selecting option *k* (based on model predictions) should closely match the observed frequency in a particular sample. The match between the observed and expected frequencies can be evaluated using chi-square fit statistics for both single items and pairs of items to investigate model fit (Drasgow, Levine, Tsien, Williams, & Mead, 1995).

However, it is well-known that chi-square statistics are affected by sample size. Thus, in large sample studies, the chi-square values will be particularly large and could result in a number items being incorrectly identified as nonequivalent. Therefore, the chi-square fit statistic should be adjusted to a smaller sample size (Chernyshenko et al., 2007) using the following formula (F. Drasgow, personal communication, March 7, 2014):

$$Adjusted\ \chi^2 = df + N_a \frac{(\chi^2 - df)}{N}$$

where the $\chi^2$ is based on the IRT model estimated, df is the degrees of freedom in the IRT model, N is the actual sample size, and $N_a$ is the adjusted sample size or the sample size that you want to calculate the chi-square for (e.g., 500). After this adjustment has been made, $\chi^2$/df ratios greater than 3 will suggest misfit (Chernyshenko et al., 2007).

As a supplement to the chi-square statistics, Drasgow et al. (1995) also suggested examining fit plots that compare the ICC from the observed data to the ICC defined by the IRT model. Again, if the model fits the data well, observed and expected ICCs should be similar. Both the adjusted chi-squares and the fit plots can be estimated in the MODFIT (Stark, 2001) or FORSCORE (Williams & Levine, 1993) computer programs.

*CFA MACS Analyses*

In contrast to the non-linear ICCs shown in Figures A1 and A2, the traditional CFA model defines a linear relationship between the latent trait and item responses. An example of this is shown in Figure A3. The x-axis shows the range of the latent trait while the y-axis represents the mean-predicted response to this particular item. This function is described by

$$\hat{X}_i = \tau_i + \lambda_i \xi$$

where $\hat{X}_i$ is the mean predicted response to item $i$, $\tau_i$ is the intercept for this item, and $\lambda_i$ is the item's loading on the latent trait. Here, $\xi$ is the individual's standing on the latent trait (similar to θ in IRT). For additional information about the assumptions that are made in the CFA model or with the maximum likelihood (ML) estimator that is commonly used to estimate these models, interested readers are referred to Bollen (1989) or Kline (2011).

In practice, tests of measurement equivalence using a MACS approach include fitting confirmatory factor models with increasingly severe restrictions on three parameters over time and/or across groups: (a) factor loadings, (b) intercepts/thresholds (continuous/categorical variables), and (c) residual variances. However, several authors have suggested that requiring equivalent error variances is overly stringent and inappropriate for many practical scenarios (Bentler, 1995; Joreskog, 1971; Vandenberg, 2002) and this view seems to be generally accepted

(Byrne, 1998). Therefore, invariance testing is typically comprised of sequentially testing three separate CFA models: configural-, metric-, and scalar-invariance models.

In the first and least restrictive model (configural invariance), the pattern of zero and non-zero loadings is tested by constraining the manifest indicators (items) to load on the same factor across age groups. Stated differently, tests of configural invariance examine whether the factor structure is the same in the reference and focal groups. Substantively, an invariant configural model indicates that each group is using the same frame of reference when responding to the survey. If this is not the case, then group-level comparisons will not be meaningful and comparing means will be equivalent to comparing apples and oranges. A failure to establish equivalent factor structures across age groups indicates that no further tests of invariance are applicable. In contrast, if configural invariance is found, a test for metric equivalence would follow and the configural model would be used as a baseline for comparing the fit of this more constrained model.

Next, the factor loadings are constrained to be equal across groups. Metric or weak invariance tests whether the same indicators (manifest variables) relate to constructs (latent variables) in the same way in each group. Given the interpretation of a factor loading in CFA models, confirmation of metric equivalence indicates that differences can be compared and allows for meaningful comparisons of factor variances and covariances across groups.

Finally, tests of scalar or strong invariance constrain both the intercepts and the factor loadings of each of the items to equivalence. When scalar invariance holds, mean-level comparisons across groups will be justified. In other words, scalar invariance suggests that the response options have the same psychological meaning across groups or at each time point in a longitudinal study.

With MACS analyses, DIF has traditionally been identified using chi-square difference tests between constrained and unconstrained nested models. However, due to the well-known sensitivity of the chi-square to sample size, some authors have suggested evaluating differences in the CFI (ΔCFI) across nested models (Cheung & Renvold, 2002; Meade et al., 2008) and simulation research has demonstrated that a ΔCFI > .002 can be an accurate indicator of nonequivalence (Meade et al., 2008).

Once appropriate models are fit to the data, both the IRT and CFA approaches can differentiate between impact (i.e., true differences in the latent trait) and measurement bias. This is the case because both methodologies model the latent trait being evaluated ($\theta$ and $\xi$ in IRT and CFA, respectively) and, as a result, can be used to test for differences in the response process across groups by controlling for the level of the latent construct. This approach is shown in Figures A1 and A3. For example, if we pick a value of the latent trait on the x-axes, measurement invariance is illustrated by differences between the predicted responses[14] across each group. Note that both the IRT and CFA approaches focus on identifying DIF by examining parameter differences across groups. When controlling for the level of the latent trait, differences between predicted responses can only result when the parameters vary across groups. In this way, impact is controlled while measurement bias is examined.

This ability to differentiate bias and impact (and the formal equivalence of these two models; Takane & de Leeuw, 1987) means that both IRT and CFA can be effective at detecting DIF under most conditions and when a consistent strategy for detecting DIF is used. Stark et al. (2006) conducted a simulation study comparing these two methodologies and noted that the

[14] The same principle applies to the SGR model and the ICC illustrated in Figure A2. However, the situation is more complicated in that the differences in the predicted responses for each category are evaluated across groups. In this case, DIF is the aggregate of differences in the option response functions and it is possible for differences in one option to cancel out differences in the opposite direction on another option. This is similar to the concept of DTF when aggregating item-level differences to assess overall functioning at the test level.

traditional approaches to detecting DIF with IRT and CFA used different strategies for identifying nonequivalence. For example, IRT researchers typically tested for DIF by constraining both the discrimination and difficulty parameters to be equal across groups at the same time. In contrast, researchers using the CFA approach generally tested for differences in the loadings and intercepts sequentially. However, when a consistent strategy for DIF detection was used, Stark et al. found that both IRT and CFA provided consistent results when used to analyze the same simulated data sets. Therefore, the results for these two methods should be similar when a consistent strategy is used to analyze the same large-scale dataset.

*Equating Parameter Estimates*

In both MACS and IRT DIF analyses, the referent items must be equivalent across groups in order to justify comparisons. Nonequivalent referent items will result in item parameters that are scaled differently in each of the samples and these differences can either exacerbate or mask true parameter differences between the groups. In other words, both IRT and CFA DIF analyses assume the equivalence of the referent items across groups and this assumption must be tested prior to conducting the analyses. Therefore, we used the approach suggested by Stark, Chernyshenko, and Drasgow (2006) to identify appropriate referent items in both IRT and CFA analyses. Based on their simulations, Stark et al. proposed that appropriate referent items could be identified using the constrained baseline approach to testing for DIF. With this approach, the parameters for all of the items in the scale are constrained to be equivalent across groups. Next, the parameters for each item are sequentially allowed to vary across groups and differences in the fit statistics are used to identify non-equivalence. Stark et al. showed that this method resulted in high Type I error rates but also had high power for detecting non-equivalence. Thus, although some items might be erroneously flagged as non-equivalent

(i.e., Type I errors), this technique is likely to detect nonequivalence if it exists. Conversely, if an item demonstrates equivalence using this technique it is highly likely that it is an appropriate item to use as the referent item. Therefore, we used this approach to identify an equivalent referent item. Once identified, the factor loading for the referent item was constrained to 1.00 and the intercept was constrained to zero in the CFA models to set the variance and mean of the latent factor, respectively.

## Appendix References

Bentler, P. M. (1995). *EQS: Structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Byrne, B. M. (1998). *Structural equation modeling in LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahweh, NJ: Lawrence Erlbaum Associates.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*, 88-106.

Cheung, G. W., & Rensvold, R. B. (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233−255.

Drasgow, F., Levine M. V., Tsien, S., Williams B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409−426.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford Press.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*, *40*, 261-279.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568−592.

Stark, S. (2001). *MODFIT: A computer program for model-data fit* [computer program]. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292−1306.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139-158.

Williams, B., & Levine, M. V. (1993). *FORSCORE: A computer program for nonparametric item response theory*. Unpublished manuscript.

Table A1

*Measurement Equivalence Results for the BFI Scales Using MACS Analyses in the Sample of 15726*

| Big 5 Factor | $\chi^2$ (df) | RMSEA | CFI | NNFI | $d_{MACS}$ |
|---|---|---|---|---|---|
| **Extraversion** | | | | | |
| 2-Factor (Age 20) | 449.97 (17) | .04 | .991 | .99 | |
| 2-Factor (Age 50) | 955.15 (17) | .06 | .978 | .96 | |
| **Configural Invariance** | 1403.13 (34) | .05 | .985 | .98 | |
| **Scalar Invariance** | | | | | |
| ...Is talkative | 1449.34* (36) | .05 | .984 | .98 | .11 |
| ...Is full of energy[a] | -- | -- | -- | -- | -- |
| ...Generates a lot of enthusiasm | 1706.78* (36) | .05 | .981** | .97 | .21 |
| ...Has an assertive personality | 1747.34* (36) | .06 | .981** | .97 | .26 |
| ...Is outgoing, sociable | 1443.17* (36) | .05 | .984 | .98 | .11 |
| ...Is reserved[a] | -- | -- | -- | -- | -- |
| ...Tends to be quiet | 1559.29* (36) | .05 | .983 | .97 | .18 |
| ...Is sometimes shy, inhibited | 1634.45* (36) | .05 | .982** | .97 | .18 |
| **Agreeableness** | | | | | |
| 2-Factor (Age 20) | 1189.00 (26) | .05 | .953 | .94 | |
| 2-Factor (Age 50) | 1189.63 (26) | .05 | .958 | .94 | |
| **Configural Invariance** | 2378.67 (52) | .05 | .956 | .94 | |
| **Scalar Invariance** | | | | | |
| ...Tends to find fault with others | 2446.70* (54) | .05 | .954 | .94 | .13 |
| ...Is helpful and unselfish with others | 2426.64* (54) | .05 | .955 | .94 | .06 |
| ...Starts quarrels with others | 2519.70* (54) | .05 | .953** | .94 | .15 |
| ...Has a forgiving nature | 2388.37* (54) | .05 | .955 | .94 | .04 |
| ...Is generally trusting | 2406.95* (54) | .05 | .955 | .94 | .08 |
| ...Can be cold and aloof[a] | -- | -- | -- | -- | -- |
| ...Is considerate and kind to almost everyone[a] | -- | -- | -- | -- | -- |
| ...Is sometimes rude to others | 2426.97* (54) | .05 | .955 | .94 | .12 |
| ...Likes to cooperate with others | 2465.70* (54) | .05 | .954 | .94 | .11 |
| **Conscientiousness** | | | | | |
| 2-Factor (Age 20) | 855.10 (26) | .05 | .976 | .97 | |
| 2-Factor (Age 50) | 1130.85 (26) | .05 | .969 | .96 | |
| **Configural Invariance** | 1986.46 (52) | .05 | .973 | .96 | |
| **Scalar Invariance** | | | | | |
| ...Does a thorough job | 2071.60* (54) | .05 | .971 | .96 | .15 |
| ...Can be somewhat careless[a] | -- | -- | -- | -- | -- |
| ...Is a reliable worker | 2122.76* (54) | .05 | .971 | .96 | .14 |
| ...Tends to be disorganized | 2205.13* (54) | .05 | .969** | .96 | .21 |
| ...Tends to be lazy | 2115.43* (54) | .05 | .971 | .96 | .16 |
| ...Perseveres until the task is finished[a] | -- | -- | -- | -- | -- |

| | | | | | |
|---|---|---|---|---|---|
| ...Does things efficiently | 2255.25* (54) | .05 | .969** | .96 | .20 |
| ...Makes plans and follows through with them | 2021.86* (54) | .05 | .972 | .96 | .07 |
| ...Is easily distracted | 2093.98* (54) | .05 | .971 | .96 | .17 |
| **Neuroticism** | | | | | |
| 2-Factor (Age 20) | 873.63 (16) | .06 | .976 | .96 | |
| 2-Factor (Age 50) | 952.42 (16) | .06 | .976 | .96 | |
| **Configural Invariance** | 1828.91 (32) | .06 | .976 | .96 | |
| **Scalar Invariance** | | | | | |
| ...Is depressed, blue[a] | -- | -- | -- | -- | -- |
| ...Can be tense | 1883.34* (34) | .06 | .975 | .96 | .11 |
| ...Worries a lot | 1861.31* (34) | .06 | .975 | .96 | .09 |
| ...Can be moody | 2120.32* (34) | .06 | .972** | .95 | .22 |
| ...Gets nervous easily | 2042.97* (34) | .06 | .973** | .96 | .22 |
| ...Is relaxed, handles stress well | 2095.63* (34) | .06 | .972** | .95 | .20 |
| ...Is emotionally stable, not easily upset | 1854.28* (34) | .06 | .975 | .96 | .06 |
| ...Remains calm in tense situations[a] | -- | -- | -- | -- | -- |
| **Openness** | | | | | |
| 2-Factor (Age 20) | 744.23 (33) | .04 | .979 | .97 | |
| 2-Factor (Age 50) | 951.45 (33) | .04 | .979 | .97 | |
| **Configural Invariance** | 1695.48 (66) | .04 | .979 | .97 | |
| **Scalar Invariance** | | | | | |
| ...Is original comes up with new ideas[a] | -- | -- | -- | -- | -- |
| ...Is curious about many different things | 1706.70* (68) | .04 | .979 | .97 | .04 |
| ...Is ingenious, a deep thinker | 1741.06* (68) | .04 | .978 | .97 | .09 |
| ...Has an active imagination | 1989.67* (68) | .04 | .975** | .97 | .22 |
| ...Is inventive | 1897.79* (68) | .04 | .976** | .97 | .15 |
| ...Prefers work that is routine (R) | 2042.70* (68) | .04 | .974** | .97 | .23 |
| ...Likes to reflect, play with ideas | 1738.65* (68) | .04 | .978 | .97 | .10 |
| ...Values artistic, aesthetic experiences[a] | -- | -- | -- | -- | -- |
| ...Has few artistic interests (R) | 1803.30* (68) | .04 | .977 | .97 | .13 |
| ...Is sophisticated in art, music, or literature | 2153.26* (68) | .04 | .973** | .96 | .22 |

Notes: *p < .05, **ΔCFI > .002.
[a] Referent items

Figure A1

*IRT Item Characteristic Curves for Two Groups*

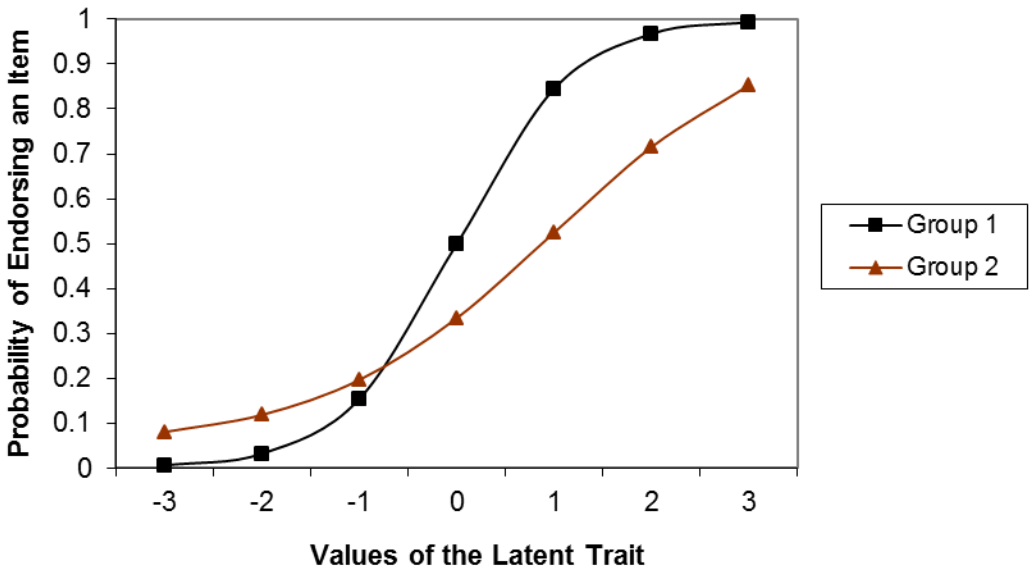


Figure A2

*Example ICCs for the SGR Model*

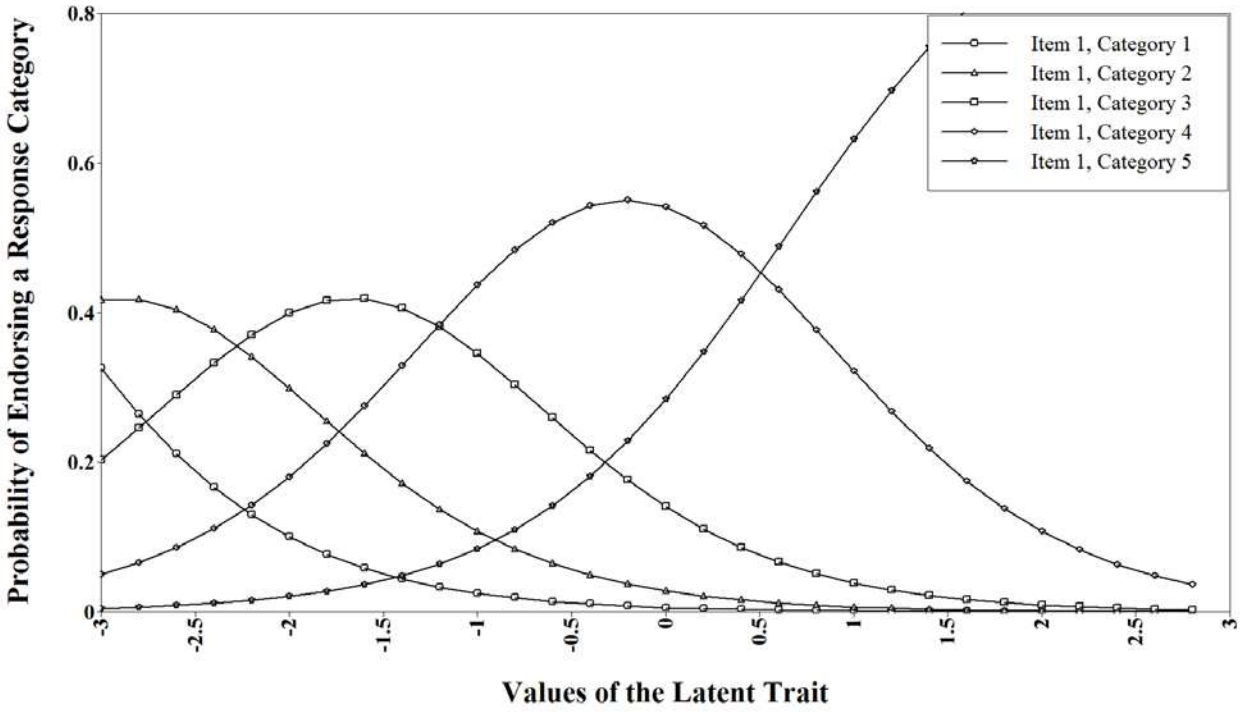
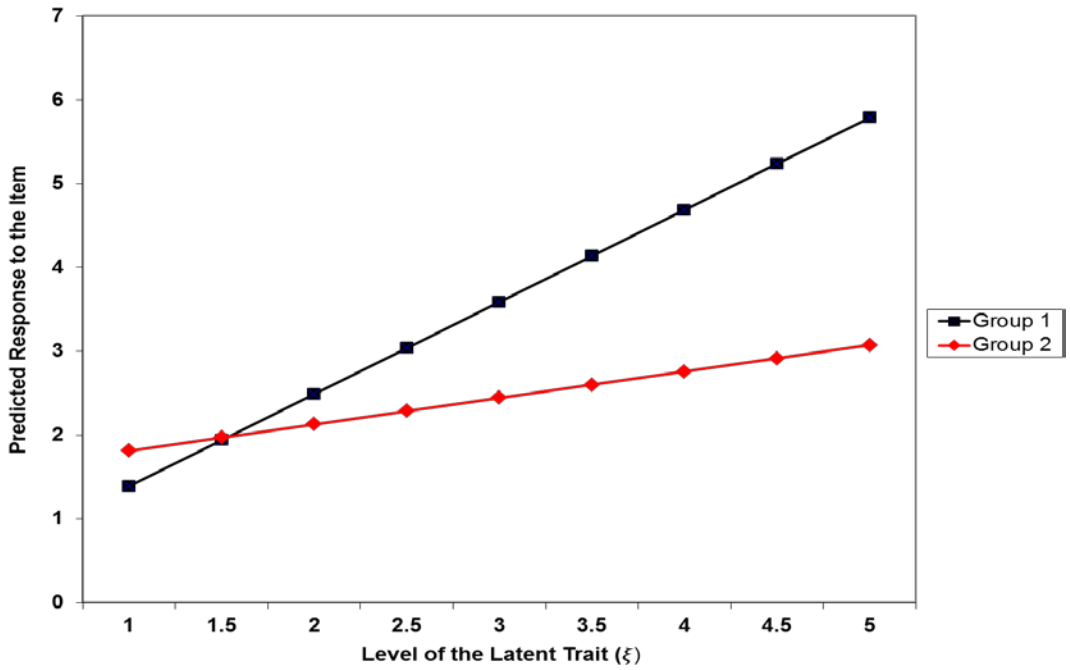


Figure A3

*Mean Predicted Responses Based on the CFA Model for Two Groups*



## Supplementary Materials

**Example Mplus syntax for the CFA scalar invariance model of the Conscientiousness scale with a single item (Csbfi3) constrained to be equivalent across groups**

```
TITLE:    Scalar Model of Conscientiousness
DATA:     FILE IS 'C:\Conscientiousness_data.dat';
          FORMAT is 1F7.0,10F1.0;

VARIABLE: NAMES ARE
recordid
Csbfi3
Csbfi13
Csbfi28
Csbfi33
Csbfi38
Csbfi8R
Csbfi18R
Csbfi23R
Csbfi43R
Group;

USEVARIABLES ARE
Csbfi3
Csbfi13
Csbfi28
Csbfi33
Csbfi38
Csbfi8R
Csbfi18R
Csbfi23R
Csbfi43R;

      MISSING = BLANK;
      GROUPING IS Group(1 = Age20 2 = Age50);

ANALYSIS:   ITERATIONS=10000;

MODEL:
       Positive BY Csbfi28@1 Csbfi3 Csbfi13 Csbfi33 Csbfi38;
       Negative BY Csbfi8R@1 Csbfi18R Csbfi23R Csbfi43R;

       [Csbfi28@0]
       [Csbfi3]
       [Csbfi13]
       [Csbfi33]
```

```
        [Csbfi38];
        [Csbfi8R@0]
        [Csbfi18R]
        [Csbfi23R]
        [Csbfi43R];

        [Positive];
        [Negative];

  MODEL Age20:
!  Csbfi3 omitted to constrain parameters to be equal across groups

        Positive BY Csbfi28@1 Csbfi13 Csbfi33 Csbfi38;
        Negative BY Csbfi8R@1 Csbfi18R Csbfi23R Csbfi43R;

        [Csbfi28@0]
        [Csbfi13]
        [Csbfi33]
        [Csbfi38];
        [Csbfi8R@0]
        [Csbfi18R]
        [Csbfi23R]
        [Csbfi43R];

        [Positive];
        [Negative];

  MODEL Age50:
!  Csbfi3 omitted to constrain parameters to be equal across groups

        Positive BY Csbfi28@1 Csbfi13 Csbfi33 Csbfi38;
        Negative BY Csbfi8R@1 Csbfi18R Csbfi23R Csbfi43R;

        [Csbfi28@0]
        [Csbfi13]
        [Csbfi33]
        [Csbfi38];
        [Csbfi8R@0]
        [Csbfi18R]
        [Csbfi23R]
        [Csbfi43R];

        [Positive];
        [Negative];

OUTPUT:  TECH1;
```

**Example Mplus syntax for the IRT DIF model of the Conscientiousness scale with a single item (Csbfi3) constrained to be equivalent across groups**

```
TITLE:    DIF Model of Conscientiousness
DATA:     FILE IS 'C:\Conscientiousness_data.dat';
          FORMAT is 1F7.0,10F1.0;

VARIABLE: NAMES ARE
recordid
Csbfi3
Csbfi13
Csbfi28
Csbfi33
Csbfi38
Csbfi8R
Csbfi18R
Csbfi23R
Csbfi43R
Group;

USEVARIABLES ARE
Csbfi3
Csbfi13
Csbfi28
Csbfi33
Csbfi38
Csbfi8R
Csbfi18R
Csbfi23R
Csbfi43R;

CATEGORICAL ARE
Csbfi3
Csbfi13
Csbfi28
Csbfi33
Csbfi38
Csbfi8R
Csbfi18R
Csbfi23R
Csbfi43R;


      MISSING = BLANK;
      CLASSES = Groups (2);
```

```
        KNOWNCLASS = Groups(Group=1 Group=2);

ANALYSIS: TYPE = MIXTURE;
                ALGORITHM = INTEGRATION;

MODEL:

%OVERALL%

        Positive BY Csbfi3* Csbfi13 Csbfi28 Csbfi33 Csbfi38;
        Negative BY Csbfi8R* Csbfi18R Csbfi23R Csbfi43R;

        Positive@1;
        Negative@1;

%Groups#1%

! Csbfi3 omitted to constrain parameters to be equal across groups
! Referent items also omitted to constrain parameters across groups

        Positive BY Csbfi13 Csbfi33 Csbfi38;
        Negative BY Csbfi18R Csbfi23R Csbfi43R;

        [Csbfi13$1-Csbfi13$4];
        [Csbfi33$1-Csbfi33$4];
        [Csbfi38$1-Csbfi38$4];

        [Csbfi18R$1-Csbfi18R$4];
        [Csbfi23R$1-Csbfi23R$4];
        [Csbfi43R$1-Csbfi43R$4];

%Groups#2%

! Csbfi3 omitted to constrain parameters to be equal across groups
! Referent items also omitted to constrain parameters across groups

        Positive BY Csbfi13 Csbfi33 Csbfi38;
        Negative BY Csbfi18R Csbfi23R Csbfi43R;

        [Csbfi13$1-Csbfi13$4];
        [Csbfi33$1-Csbfi33$4];
        [Csbfi38$1-Csbfi38$4];

        [Csbfi18R$1-Csbfi18R$4];
        [Csbfi23R$1-Csbfi23R$4];
        [Csbfi43R$1-Csbfi43R$4];
```

```
        [Csbfi23R$1-Csbfi23R$4];
        [Csbfi43R$1-Csbfi43R$4];


OUTPUT:  TECH1 TECH8;
PLOT:    TYPE = PLOT3;
```